December 13, 2023

*E-Filed*

The Honorable Donna M. Ryu
United States District Court for the Northern District of California
Oakland Courthouse, Courtroom 4 – 3rd Floor
1301 Clay Street
Oakland, California 94612

Re: *Doe 1, et al. v. GitHub, Inc., et al.;* Lead Case No. 4:22-cv-06823-JST

Dear Judge Ryu:

Pursuant to the Court's Order Referring All Discovery Disputes to Magistrate Judge (ECF No. 175) and Paragraph 14 of this Court's Standing Order ("Standing Order"), Plaintiffs Does 1–5 ("Plaintiffs") submit this letter brief regarding GitHub, Inc.'s ("GitHub") responses to Plaintiffs' First Set of Interrogatories propounded on July 19, 2023 ("Interrogatories"). Plaintiffs seek an order compelling GitHub to provide responses to Interrogatories 2, 6, 7, and 14.

Consistent with the Standing Order, Plaintiffs attempted to further meet and confer with GitHub regarding its plainly inadequate discovery responses. GitHub has refused to meet and confer in person. Plaintiffs have previously met and conferred by zoom on August 29 and September 7, 2023. Plaintiffs subsequently confirmed the parties' positions in writing on September 19, 2023. Even after Your Honor's appointment, GitHub has dragged its feet and avoided its discovery obligations. Plaintiffs have made efforts to submit a joint letter. GitHub has refused. Plaintiffs provided a copy of this letter to GitHub prior to filing. GitHub's approach is inconsistent with Rules 26, 33, 37, as well as the local rules of the Court and the Standing Order.

**Current Case Deadlines:** (1) Fact Discovery Cut-Off: September 27, 2024; (2) Expert Discovery Cut-Off: February 21, 2025; (3) Dispositive Motion Hearing: not scheduled; (4) Class Certification Motion: March 27, 2025; (4) Pretrial Conference and Trial Dates: not scheduled.

**Interrogatory 2** calls for identification of persons with relevant information, specifically the identities of persons who have managed and directed GitHub during the relevant period. This is basic information, routinely provided in civil litigation. *See, e.g.*, *In re TFT-LCD (Flat Panel) Antitrust Litig.*, No. M 07-1827 SI, 2007 WL 2782951, at *2 (N.D. Cal. Sept. 25, 2007) (permitting interrogatories seeking the "names, positions, dates of employment/tenure, and addresses" of *inter alia*, directors and officers as an exception to a general discovery stay); *In re Folding Carton Antitrust Litig.*, 76 F.R.D. 417, 419 (N.D. Ill. 1977) (describing this information as "classic first-wave discovery"). GitHub states it "will not undertake the investigation necessary to respond." Even though GitHub has never offered to produce any responsive information, Plaintiffs agree to narrow Interrogatory 2 to request names and dates in such positions of past and present officers and directors in one of those positions from 2017 on, including the period Microsoft acquired GitHub.

**Interrogatory 6** seeks names of individuals responsible for negotiating GitHub's 2018 acquisition by Microsoft for $7.5 billion. Microsoft and GitHub had close connections pre-acquisition.[1] GitHub's library of open-source code was used to train Copilot. Microsoft purchased GitHub's

[1] *See* https://www.theverge.com/2018/6/4/17422788/microsoft-github-acquisition-official-deal/.

library, including the repositories containing the code subject to the licenses at issue in the case, related products, and product conceptualizations. Responsive individuals—including GitHub's agents—will have relevant knowledge, including plans for products that became Copilot. They will know the basis for the $7.5 billion purchase price and can also reasonably be expected to know what, if any, plans or actions were taken regarding the open-source licenses such as whether and how to ignore or violate them. This interrogatory merely seeks witness identities. Rule 33 requires the responding party to provide information using any information readily available to the responding party, including information known by employees, or any information contained in files maintained, or otherwise available.

**Interrogatory No. 7**—like Interrogatory 6—seeks straightforward information about GitHub's purchases and/or sales of an ownership interest in OpenAI. The nature of the business relationship between GitHub and OpenAI, the co-creators of Copilot, is a core factual issue in this case. This information will shed light on how the training model was developed and chosen, the economic terms, as well as GitHub's knowledge, scienter, and affirmative defenses. The transactions are not public, and not available from other sources. As with Interrogatory 6, the burden of responding here is low. *See De Vera v. United Airlines Inc.*, No. C 12-05644 LB, 2013 WL 12182141, at *1 (N.D. Cal. Sept. 9, 2013). GitHub's description of the burden it ostensibly faces is that it would have to "investigate whether any of its employees at any time sold or purchased any interest in OpenAI on their own behalf." This assertion of burden is speculative and inadequate. OpenAI is a private company, and the number of GitHub entities or even personnel owning stock is likely small and concentrated. Even if responding would create some burden, the fact that OpenAI has not even attempted to collect the necessary information is contrary to its discovery obligations under Rule 26. *In re ATM Fee Antitrust Litig.*, 233 F.R.D. 542, 545 (N.D. Cal. 2005). Presumably, the company maintains a capitalization table or other similar document where such information is maintained and could easily be produced. GitHub has not offered any response.

**Interrogatory No. 14** seeks information regarding the identities of individuals that participated in acquiring or licensing training data at issue in this case and details of the terms. The extent of GitHub's rights to the data used for training Copilot, how it was obtained, and details of its CMI are core issues here. *See, e.g.*, First Amended Complaint (ECF No. 135) ("FAC") ¶¶ 2, 10, 84–95, 128. Courts routinely permit discovery necessary to ascertain the extent of copyright-related violations. *See, e.g.*, *Transglobal Commc'ns Grp., USA, Inc. v. Stone Sapphire, Ltd.*, No. CV 07-0500-GW(RCX), 2008 WL 11339591, at *2 (C.D. Cal. Mar. 21, 2008) (ordering defendants to respond to interrogatory which asked for information about factories defendants have used or use, which "relates to plaintiff's allegation that defendants are manufacturing or producing products containing plaintiff's copyrighted works."). In addition, for example, Plaintiffs must establish scienter in connection with removal of CMI. Details regarding the procurement of the training data and its CMI are essential support for this claim. GitHub does not dispute it possesses such information. Remarkably, GitHub t rewrites the interrogatory to seek only "data used to train the OpenAI models underlying Copilot" and then fails to provide it. A complete description of data used to train GPT-3, GPT-4, Codex, and Copilot remains secret. Plaintiffs only know Codex was trained on "billions of lines of source code from publicly available sources, including code in public GitHub repositories." FAC ¶ 87.[2] Composition of the training data for these models will be known to responsive individuals. Plaintiffs agree to accept a complete list of all data used to train GPT-3, GPT-4, Codex, and Copilot and the identities of any individual that collected any of that data.

[2] *See also* https://github.blog/2021-06-30-github-copilot-research-recitation/.

*/s/ Joseph R. Saveri*
Joseph R. Saveri
Joseph Saveri Law Firm, LLP
Attorneys for Plaintiffs